# Can AI-estimated article quality be used to rank scholarly documents?

Mike Thelwall

*University of Wolverhampton, Wolverhampton WV1 1LY, UK*

### Abstract

This paper discusses the potential for machine learning predict the quality of scholarly documents to help rank them in information retrieval systems. Quality-based rankings may help users without the time or expertise to assess the value of the publications suggested by a system. It is argued that systems to learn the quality of documents with a degree of accuracy may be possible from the increasing availability of reviews and scores online.

A key feature of scholarly information retrieval systems is their ranking algorithms. Users may focus on the first documents that they see, unless they are attempting a comprehensive review. The use of citation information to rank scholarly search results is arguably appropriate for academics because citations are an obvious, but partial, indicator of scholarly uptake or utility. A document that has been cited a lot is very likely to have been read by many publishing researchers and found useful enough to cite. In contrast, end users may be more interested in applied research. For this goal, citations may be less helpful, especially if they tend to point to basic or methodological papers rather than practical applications. Their value may also be undermined by attempts to manipulate them (e.g., [1]). For end users, it may therefore be better to rank papers by quality rather than by citation impact. Ranking-by-quality may also help in the era of predatory publishing, by pointing end users and junior academics to high quality work that is relevant to their needs. Both user groups may lack the time or experience to perform effective quality control on search results. Whilst this issue may be resolved by collaborative filtering approaches (e.g., [2]) it would be useful to rank documents before they have been seen.

The main reason why academic articles are not ranked by quality in any mainstream scholarly database may be that such quality scores are not available for most articles. Both journals and conferences usually make binary publishing decisions (accept/reject) after reviewing and do not publish a quality assessment or reviewers' quality scores. Since there are increasingly many exceptions (e.g., some open peer review conferences, F1000 post-publication ratings [3]) and there may be a future increase in post-publication peer review scores for articles, there soon may be enough public peer review scoring data for systems to harness, when available. The score data may be supplemented by algorithms to classify reviews or post-publication comments (e.g., [4]) for sentiment, or to detect problematic content in articles (e.g., [5, 6]). An alternative

method of generating article quality scores would be to apply machine learning on a sample of articles with peer reviews, perhaps from the aforementioned sources, and then use the trained algorithms to estimate the quality scores for the remainder.

Following on from the above, is it possible and desirable to use machine learning to estimate the quality of an academic article to support ranking in academic information retrieval systems? One study has predicted proxy-quality scores for articles with machine learning, using journal impact (split into thirds) as a proxy for article quality. Using this heuristic, it is possible to generate proxy quality predictions that are substantially above the baseline in some fields, but not others [7]. This suggests that automatically detecting quality will be much more difficult in some fields than in others. Intuitively, more hierarchical fields with standardised methods would be more easily to check quality for, given that deviations from best practice could theoretically be detected. In contrast, in a humanities field, it might take substantial or wide field knowledge to judge the quality of outputs.

Preliminary unpublished experiments to predict article quality with machine learning applied to tens of thousands of human quality scores (high, medium, low) for articles in 27 Scopus broad fields suggest that the highest accuracy is possible for the following biomedical and physical science Scopus broad fields: Multidisciplinary; Biochemistry, Genetics and Molecular Biology; Physics and Astronomy; Chemistry. In contrast, this task is most difficult or impossible in the following Scopus broad fields: Engineering; Agricultural and Biological Sciences; Psychology; Social Sciences; Environmental Science; Energy; Arts and Humanities; Dentistry; Nursing; and Pharmacology & Toxicology. Thus, the potential for harnessing machine learning for article quality prediction may be restricted to the biomedical and physical sciences.

Based on previous studies on predicting citation counts [8, 9, 10, 11], the following recommendations are made for a ranking system to reflect article quality in fields where it is possible.

- Journal impact thirds, quartiles or other groupings can be used as the target of a machine learning system in fields in which journal impact is a reasonable indicator of article quality (medicine, health, physical sciences, economics, psychology) but not in areas where citations have little value (engineering, other social sciences, arts and humanities). This could be replaced by post publication or peer review scores when they become available in sufficient numbers. If this replacement is made, then a journal impact indicator could become an input.
- Machine learning should be applied to data segmented into narrow coherent fields to give the algorithms the chance to learn field-specific quality patterns.
- Inputs should be field and year normalised (e.g., not citation counts but normalised variants such as the Mean Normalised Citation Score (MNCS) or the Mean Normalised Log-transformed Citation Score (MNLCS)) so that related fields and years can be combined to gain sufficient training data.
- Valuable types of inputs include all those shown to associate with citation rates, including: (normalised) article citations, number of authors, number of institutional affiliations, article length, number of country affiliations, career publishing statistics of the authors, and abstract readability.

- Text inputs, such as words and phrases used in the article title and abstract may reveal important topics, which are more relevant to citations than quality. They may also point to high quality methods (e.g., randomised control trials) and identify more subtle indicators of high-quality work, such as appropriate hedging or shared data/code. If full text can be analysed, then factors like the number of figures and tables in a paper may be useful in judging the amount of evidence supporting the article in some fields.

The above approach is clearly quite citation-dependant but at least moves one step away from a pure reliance on citations.

# References

[1] Oriensubulitermes inanis [pseudonym], PubPeer comment https://pubpeer.com/publications/940C291607CF03969C6A936F8BA5B9#2, 2022.

[2] D. Kershaw, B. Pettit, M. Hristakeva, K. Jack, Learning to rank research articles: A case study of collaborative filtering and learning to rank in ScienceDirect, in: Proceedings BIR 2020, 2020, pp. 75–88. URL: http://ceur-ws.org/Vol-2591/paper-08.pdf.

[3] M. Thelwall, E. Papas, Z. Nyakoojo, L. Allen, V. Weigert, Identification of highly-cited papers using topic-model-based and bibliometric features, Online Information Review 44 (2020). doi:https://doi.org/10.1108/OIR-11-2019-0347.

[4] J. L. Ortega, Classification and analysis of PubPeer comments: How a web journal club is used, Journal of the Association for Information Science and Technology (2021). doi:https://doi.org/10.1002/asi.24568.

[5] G. Cabanac, C. Labbé, Prevalence of nonsensical algorithmically generated papers in the scientific literature, Journal of the Association for Information Science and Technology 72 (2021) 1461–1476.

[6] G. Cabanac, C. Labbé, A. Magazinov, Tortured phrases: A dubious writing style emerging in science. Evidence of critical issues affecting established journals, 2021. arXiv preprint arXiv:2107.06751.

[7] M. Thelwall, Can the quality of published academic journal articles be assessed with machine learning?, Quantitative Science Studies (2022). doi:https://doi.org/10.1162/qss_a_00185refs.

[8] A. Abrishami, S. Aliakbary, Predicting citation counts based on deep neural network learning techniques, Journal of Informetrics 13 (2019) 485–499.

[9] Y. H. Hu, C. T. Tai, K. E. Liu, C. F. Cai, Identification of highly-cited papers using topic-model-based and bibliometric features, Journal of Informetrics 14 (2020).

[10] J. Xu, M. Li, J. Jiang, M. Cai, Early prediction of scientific impact based on multi-bibliographic features and convolutional neural network, IEEE Access (2019) 92248–92258.

[11] T. van Dongen, G. Wenniger, L. Schomaker, SChuBERT: Scholarly document chunks with bert-encoding boost citation count prediction, 2020. arXiv preprint arXiv:2012.11740.